UNI-X: MITIGATING MODALITY CONFLICT WITH A TWO-END-SEPARATED ARCHITECTURE FOR UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified Multimodal Models (UMMs) built on shared autoregressive (AR) transformers are attractive for their architectural simplicity. However, we identify a critical limitation: when trained on multimodal inputs, modality-shared transformers suffer from severe gradient conflicts between vision and text, particularly in shallow and deep layers. We trace this issue to the fundamentally different low-level statistical properties of images and text, while noting that conflicts diminish in middle layers where representations become more abstract and semantically aligned. To overcome this challenge, we propose Uni-X, a two-end-separated, middle-shared architecture. Uni-X dedicates its initial and final layers to modality-specific processing, while maintaining shared parameters in the middle layers for high-level semantic fusion. This X-shaped design not only eliminates gradient conflicts at both ends but also further alleviates residual conflicts in the shared layers. Extensive experiments validate the effectiveness of Uni-X. Under identical training conditions, Uni-X achieves superior training efficiency compared to strong baselines. When scaled to 3B parameters with larger training data, Uni-X matches or surpasses 7B AR-based UMMs, achieving a GenEval score of 82 for image generation alongside strong performance in text and vision understanding tasks. These results establish Uni-X as a parameter-efficient and scalable foundation for future unified multimodal modeling. Our code is available at `https://anonymous.4open.science/r/Uni-X-Code-E5CD`.

1 INTRODUCTION

Vision-Language Models (VLMs) have demonstrated remarkable progress in multimodal understanding and reasoning, enabled by combining Large Language Models (LLMs) with powerful visual encoders (Liu et al., 2024c;b; Wang et al., 2024b; Team et al., 2024b). Motivated by this success, recent research has sought to extend VLMs with image generation capabilities, resulting in the development of **Unified Multimodal Models (UMMs)** (Team, 2025; Wang et al., 2024d; Wu et al., 2024b). However, many state-of-the-art UMMs rely on increasingly complex system designs to boost performance, including the addition of semantic image encoders (Wu et al., 2024a; Chen et al., 2025b; Deng et al., 2025; Wu et al., 2025a), the hybridization of autoregressive and diffusion paradigms (Wu et al., 2025a; Zhao et al., 2024; Ge et al., 2025; Deng et al., 2025; Xie et al., 2025; Zhou et al., 2024), or the introduction of task-specific branches and experts (Deng et al., 2025; Liao et al., 2025; Li et al., 2025c). While effective, this added complexity hinders scalability, limiting the degree of parameter sharing and reducing the potential for mutual benefits across tasks and modalities.

In contrast, **autoregressive (AR) UMMs** offer a simple yet powerful alternative. By treating visual inputs as a "foreign language" through vector quantization (VQ) (van den Oord et al., 2018; Esser et al., 2021b), they unify text and vision into a consistent token sequence, naturally extending the language-centric paradigm of LLMs (Wu et al., 2025b; Wang et al., 2024d). Despite this simplicity, our experiments reveal a fundamental challenge: **fully modality-shared transformers trained jointly on multimodal inputs exhibits severe gradient conflicts.** Originally studied in multi-task learning (Yu et al., 2020; Shi et al., 2023), we are the first to transfer this concept to UMMs, uncovering inter-modality conflicts that hinder convergence and performance.

As illustrated in Figure 1, these conflicts are most pronounced in the shallow (input) and deep (output) layers, where the model must reconcile the vastly different statistical properties of text and images. In contrast, the middle layers, where representations become increasingly abstract and semantic (Meng et al.; Geva et al., 2021; Sun et al., 2025), show reduced conflicts and stronger cross-modal alignment. This suggests that an effective UMM should **respect modality-specific differences** rather than enforcing uniform parameter sharing across all layers.

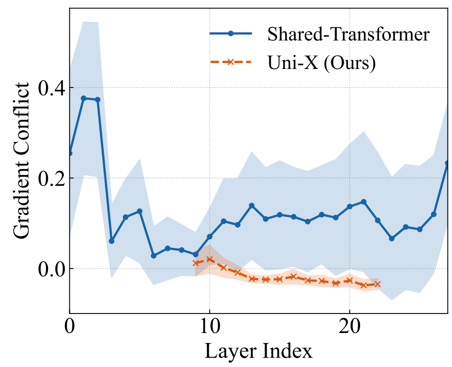

Figure 1: Gradient conflict analysis of down-projection weights in the FFN of a modality-shared transformer. The shared transformer exhibits severe conflicts in shallow and deep layers, with only partial mitigation in intermediate layers. In contrast, Uni-X avoids conflicts at both extremes and further alleviates them in the middle layers.

Guided by this observation, we introduce **Uni-X**, a *two-end-separated, middle-shared* architecture for unified multimodal modeling. In Uni-X, the shallow and deep layers are modality-specific, enabling specialized processing of distinct low-level distributions in text and vision, while the middle layers are shared to capture high-level semantic abstractions common to both. This **X-shaped architecture** not only mitigates the severe gradient conflicts at the two ends but also further alleviates residual conflicts in the shared middle layers by leveraging natural semantic alignment between modalities (Figure 1).

To demonstrate the effectiveness of Uni-X, we conduct extensive experiments under controlled training budgets and scaling regimes. Results show that Uni-X improves training efficiency and achieves stronger performance under identical conditions. Moreover, with larger data and model scales, our 3B-parameter Uni-X matches or surpasses the performance of existing 7B AR-based UMMs across both understanding and generation benchmarks, demonstrating its scalability and competitiveness. Our contributions are threefold:

- **Empirical Analysis:** We identify and quantify gradient conflicts between text and vision modalities in the shallow and deep layers of shared autoregressive transformers, attributing them to fundamental differences in their low-level statistical properties.
- **Model Design:** We propose Uni-X, a novel two-end-separated, middle-shared architecture that aligns model structure with modality characteristics by using modality-specific layers for low-level processing and a shared core for high-level semantic fusion.
- **Comprehensive Validation:** Extensive experiments demonstrate that Uni-X improves training efficiency and scales effectively, enabling a 3B model to achieve performance competitive with much larger 7B models across diverse multimodal benchmarks.

2 Related Work

Visual Language Models (VLMs). The remarkable progress of LLMs (Touvron et al., 2023; Yang et al., 2024; Brown et al., 2020) has motivated researchers to extend them with visual cognition, giving rise to VLMs (Liu et al., 2024c; Achiam et al., 2023). Most VLMs leverage pre-trained visual encoders such as CLIP (Radford et al., 2021) or SigLIP2 (Tschannen et al., 2025) to extract semantic features from images, which are projected into the LLM's semantic space via multimodal adapters (Liu et al., 2024c;b; Beyer et al., 2024; Team et al., 2024a; Li et al., 2025a). This design enables strong multimodal understanding and reasoning but remains **asymmetric**: VLMs treat images only as inputs and cannot generate them, limiting synergy between perception and synthesis.

Unified Multimodal Models (UMMs). To enable such synergy, recent efforts have shifted toward UMMs, which aim to support both understanding and generation within a single framework (Team, 2025; Jin et al., 2024; Wu et al., 2024b). A natural extension is to adopt the autoregressive (AR) paradigm of LLMs by treating visual tokens as a "foreign language" via vector quantization (Wu et al., 2025b; Wang et al., 2024c). However, the distinct statistical properties of text and images often lead to **modality conflicts**, degrading performance in shared transformers (Team, 2025).

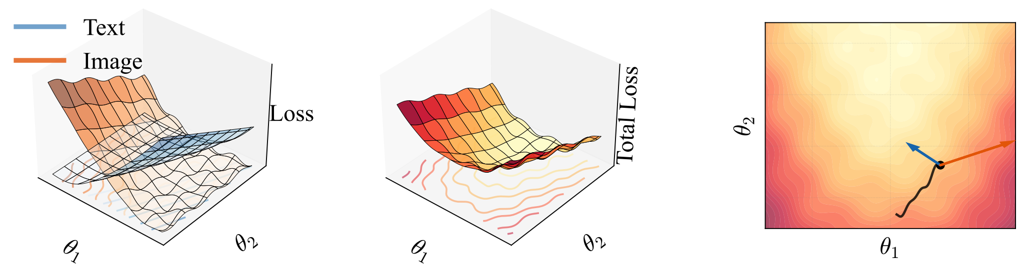

Figure 2: Illustration of gradient conflict. (a) The loss landscapes of different modalities exhibit distinct geometries, creating potential conflicts in optimization direction. (b) The optimum of the sum of losses is different from the optimum of any single modality's loss. (c) In the presence of gradient conflict, the optimization trajectory becomes oscillating and suffers from slow convergence.

To mitigate this, several approaches increase architectural complexity: Mixture-of-Transformers (MoT) designs (Liao et al., 2025; Deng et al., 2025; Shi et al., 2025) separate understanding and generation with distinct branches; Hybrid AR–diffusion frameworks (Zhao et al., 2024; Wu et al., 2025a; Dong et al., 2023; Ge et al., 2025) combine next-token prediction with diffusion-based image synthesis; and branching strategies such as UniFork (Li et al., 2025c) add task-specific deep heads. While effective on benchmarks, these methods sacrifice parameter sharing, complicate training, and weaken cross-modal benefits–the very goals UMMs were meant to unify.

Comparison with Uni-X. Uni-X builds on these insights but takes a different path. Instead of adding modules, Uni-X retains the simplicity of pure AR UMMs while mitigating modality conflict through a **two-end-separated, middle-shared** architecture: shallow and deep layers use modality-specific parameters to process low-level statistical differences, while intermediate layers are shared to exploit high-level semantic alignment. This X-shaped design avoids the rigidity of MoT or UniFork and the complexity of AR-diffusion hybrids, offering a lightweight yet effective solution. Empirically, Uni-X achieves comparable or superior performance to larger 7B AR UMMs across text, vision understanding, and generation tasks, while maintaining parameter efficiency and training simplicity.

3 Uni-X

3.1 Observations

Before introducing the Uni-X architecture, we first analyze the gradient conflicts that arise when training modality-shared autoregressive transformers on multimodal data. We also provide an information-theoretic perspective to explain why such conflicts emerge.

Definition of Gradient Conflict. As illustrated in Figure 2, gradient conflict occurs when different optimization objectives induce gradients pointing in divergent directions, making joint optimization unstable and inefficient. To quantify gradient conflict in multimodal training, we use an early checkpoint of a fully shared transformer. From this model checkpoint, we compute the *average gradients* for specific parameter groups (e.g., FFN down-projection weights) using over 60 mini-batches and totaling 2M tokens, to ensure obtaining stable gradients.

Specifically, we first compute the average text gradient, $\boldsymbol{g}_{\texttt{text}}$. This is obtained by exclusively performing forward and backward passes on $D_{\texttt{text}}$, a text-only subset filtered from the pre-training data. Next, we compute the average image-text gradient, $\boldsymbol{g}_{\texttt{img}}$, using an analogous subset $D_{\texttt{img}}$ containing image-text pairs. The raw inter-modal similarity is then measured as the cosine similarity between these two average gradients:

$$S_{\texttt{inter}} = \cos(\boldsymbol{g}_{\texttt{text}}, \boldsymbol{g}_{\texttt{img}}). \quad (1)$$

However, since transformer layers have inherently different roles across depth (Sun et al., 2025; Geva et al., 2021), the resulting raw similarity $S_{\texttt{inter}}$ is biased and cannot be directly compared. To correct for this, we estimate a baseline similarity $S_{\texttt{base}}$ that reflects the model's intrinsic gradient consistency on a unified data distribution. We randomly shuffle the full multimodal dataset $D_{\texttt{all}}$ and split it into two disjoint halves, $D^1_{\texttt{any}}$ and $D^2_{\texttt{any}}$. Their respective average gradients, $\boldsymbol{g}^1_{\texttt{any}}$ and $\boldsymbol{g}^2_{\texttt{any}}$, yield:

$$S_{\texttt{base}} = \cos(\boldsymbol{g}^1_{\texttt{any}}, \boldsymbol{g}^2_{\texttt{any}}). \tag{2}$$

This value represents the expected gradient similarity when gradients originate from the same underlying distribution. Therefore, we define the **gradient conflict** c_g as the deviation from this baseline:

$$c_g = -(S_{\texttt{inter}} - S_{\texttt{base}}). \tag{3}$$

A high $S_{\texttt{base}}$ indicates the model's gradients are stable, whereas a much lower $S_{\texttt{inter}}$ suggests that the text-only and image-text data push the shared model parameters in conflicting directions, resulting in a large positive c_g. This provides a principled, layer-wise measure of inter-modal *disagreement* and reveals where and why conflicts are most severe.

Empirical Findings. Figure 1 shows gradient conflict profiles (c_g) across depth. In modality-shared transformers, conflicts are most pronounced in shallow layers (near input) and deep layers (near output), while intermediate layers exhibit weaker conflicts. Experiments further reveal that Uni-X avoids conflicts at both extremes and reduces residual conflicts in the middle, validating its structural design. Additional analyses of other modules and the relationship between gradient conflict and data, as well as its impact on model performance, are provided in Appendix A.4.

Why Do Conflicts Arise? Vision as a "Foreign Language" To explain these observations, we examine whether vision behaves like a "foreign language" when tokenized. Using the VQ tokenizer (Team, 2025), images are represented as discrete token sequences, formally similar to text. Then, we define conditional entropy based on n-gram. When $n = 1$, the calculation reduces to ordinary information entropy. For $n > 1$, the conditional entropy is computed as follows:

$$H_n = -\sum p(w_n \mid w_1, w_2, \cdots, w_{n-1}) \log p(w_n \mid w_1, w_2, \cdots, w_{n-1}). \tag{4}$$

Results (Figure 3) show that image tokens exhibit far higher entropy than natural languages such as English, German, or Chinese. While languages differ in grammar and lexicon, their token statistics remain closer to each other than to images. This means visual sequences are inherently harder to predict, requiring modeling of long-range, spatially entangled dependencies.

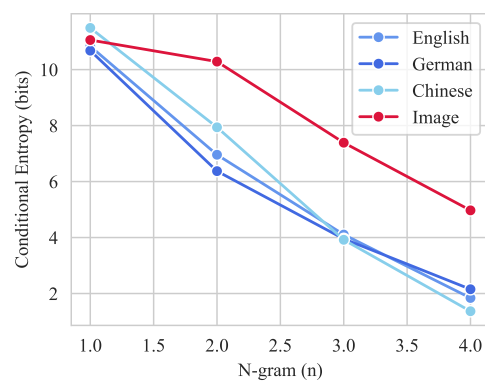

Figure 3: Conditional entropy of images and natural languages. Image token sequences encoded by the VQ tokenizer exhibit substantially higher entropy, indicating greater difficulty in prediction.

As information theory suggests, sequences with higher (conditional) entropy are inherently harder to predict, requiring models to learn longer-range dependencies and more complex patterns. Thus, when a shared transformer jointly processes low-entropy, grammatical text with high-entropy, spatially complex vision, shallow and deep layers are forced to reconcile conflicting low-level distributions, producing strong gradient conflicts. In contrast, intermediate layers, where representations become more abstract and semantic, naturally align across modalities, explaining the reduced conflicts observed in practice.

3.2 Model Architecture

Motivated by these findings, we propose **Uni-X**, an architecture designed to explicitly align model structure with modality characteristics.

Core Principle. As illustrated in Figure 4, Uni-X follows a two-end-separated, middle-shared design. The shallow and deep layers are duplicated into parallel modality-specific branches, ensuring independent handling of text and vision during early feature extraction and final token projection.

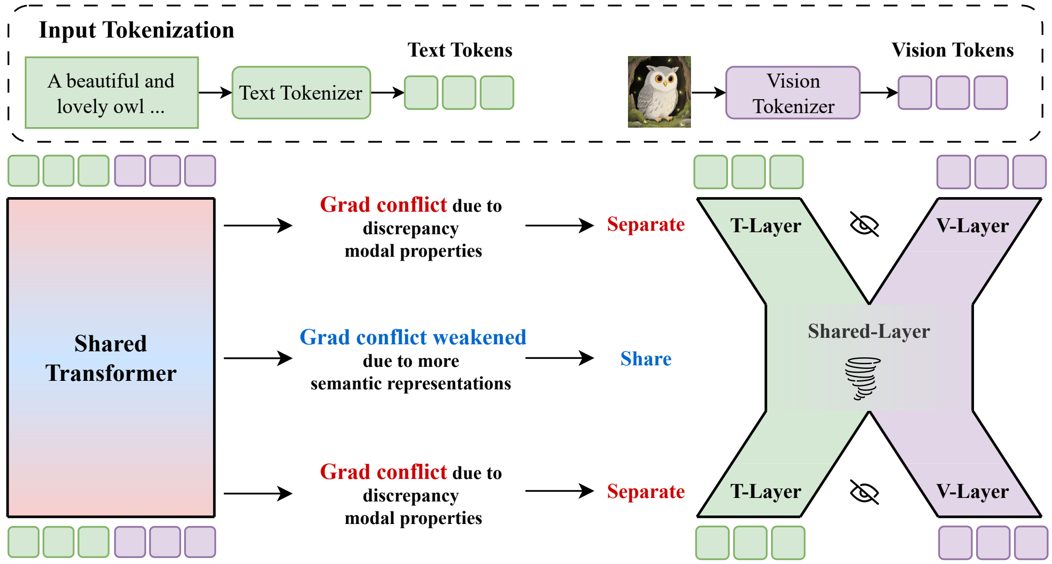

Figure 4: Illustration of the proposed **Uni-X** architecture compared with a standard modality-shared transformer. The baseline shared transformer (**left**) encounters gradient conflicts in shallow and deep layers due to the mismatched statistical properties of vision and text tokens. In contrast, Uni-X (**right**) adopts a two-end-separated, middle-shared design: modality-specific layers at both ends handle low-level feature processing, while a shared central block performs high-level semantic fusion. This structure aligns the architecture with the inherent characteristics of each modality and effectively mitigates gradient conflicts.

The intermediate layers remain shared, enabling high-level semantic fusion across modalities. This X-shaped separation-and-sharing balances modality specialization with semantic alignment.

Input Tokenization. For visual inputs, we employ the VQGAN tokenizer (Esser et al., 2021a) from Chameleon (Team, 2025) to encode 512×512 images into a 32×32 grid of visual tokens from an 8,192-entry codebook. To accommodate these new tokens, we expand the vocabulary and corresponding embedding matrix of the base LLM. Textual inputs are processed via the standard BPE tokenizer. The final token sequence is structured as <BOI>[Image]<EOI>[Text]<BOS> for image understanding tasks and [Text]<BOI>[Image]<EOI><BOS> for image generation tasks. This unified tokenization enables AR training across modalities. While Uni-X can handle interleaved multimodal sequences, this study focuses on non-interleaved inputs.

Forward Propagation. Given a pre-trained LLM with L layers, denoted as $\{\texttt{Layer}_{\texttt{t}}^i\}_{i=0}^{L-1}$, we partition them into three sections: The initial N layers and the final M layers constitute the "separated layers," while the intermediate layers form the "shared layers." Within the separated blocks, we introduce a new set of vision-specific layers, $\{\texttt{Layer}_{\texttt{v}}^i\}$, which operate in parallel with the original text layers, $\{\texttt{Layer}_{\texttt{t}}^i\}$.

To manage the data flow, we introduce a binary mask $\boldsymbol{M}_{\texttt{v}} \in \{0,1\}^n$ to identify the positions of visual tokens. At any given layer l, the complete hidden states $\boldsymbol{H}^l$ can be partitioned into text-specific states $\boldsymbol{H}_{\texttt{t}}^l = \boldsymbol{H}^l[\sim \boldsymbol{M}_{\texttt{v}}]$ and vision-specific states $\boldsymbol{H}_{\texttt{v}}^l = \boldsymbol{H}^l[\boldsymbol{M}_{\texttt{v}}]$.

The forward propagation in Uni-X is defined as follows:

$$\boldsymbol{H}_x^{l+1} = \begin{cases} \texttt{Layer}_x^l(\boldsymbol{H}_x^l) & \text{if } l < N \text{ or } l \geq L - M, \\ \left[\texttt{Layer}_{\texttt{t}}^l(\boldsymbol{H}^l)\right]_x & \text{otherwise,} \end{cases} \tag{5}$$

where $x \in \{\texttt{t}, \texttt{v}\}$ denotes the modality (text or vision). In the "otherwise" case, $[\cdot]_x$ indicates selecting the subset of the output hidden states corresponding to modality x.

Importantly, unlike other architectures (Li et al., 2025c; Deng et al., 2025; Shi et al., 2025), the vision and text modalities remain strictly isolated within the separated blocks, with no cross-modal interaction. This forces the model to learn robust unimodal representations before they are fused in the shared block and after they are separated for modality-specific output generation.

Training Objective. Following the standard paradigm for AR models, Uni-X is trained to predict the next token in a sequence containing both text and visual tokens. The training objective is to minimize the cross-entropy loss over the vocabulary for each token. The loss function $\mathcal{L}$ for a given sequence $\mathcal{S} = (s_1, s_2, \cdots, s_T)$ is defined as:

$$\mathcal{L} = -\sum_{i=1}^{T} \log P(s_i \mid s_{<i}). \tag{6}$$

This simple yet effective objective enables the model to learn both understanding and generation capabilities across modalities within a single, unified framework.

Design Rationale. Unlike prior architectures that rely on auxiliary semantic encoders (Wu et al., 2024a; Chen et al., 2025b), hybrid AR–diffusion pipelines (Wu et al., 2025a; Zhao et al., 2024), or task-specific branching structures such as MoT (Liao et al., 2025) and UniFork (Li et al., 2025c), Uni-X maintains the simplicity of a pure autoregressive framework. Its two-end-separated, middle-shared structure is motivated directly by empirical evidence of gradient conflicts, aligning the model design with the statistical characteristics of each modality. By isolating low-level modality-specific processing while preserving a shared semantic core, Uni-X avoids the complexity and overhead of multi-expert or dual-paradigm systems, yet achieves competitive or superior performance. This balance of architectural simplicity, empirical grounding, and scalability makes Uni-X a practical foundation for unified multimodal modeling.

4 Experiments

We evaluate Uni-X from two complementary perspectives: (1) **Efficiency under identical training conditions**, where Uni-X and baseline architectures are trained on the same data and resources, enabling fair comparisons of efficiency and performance. (2) **Scaling within resource constraints**, where we maximize dataset size and training duration to examine Uni-X's scalability and competitiveness against larger state-of-the-art models.

4.1 Experimental Setup

Pre-training Datasets. Our pre-training stage was designed to build a strong foundation in both language and vision. To preserve general text generation capabilities, we utilized a diverse set of text corpora: the high-quality Chinese dataset CCI3-H (Wang et al., 2024a), English datasets DCLM (Li et al., 2025b) and Fineweb-Edu (Penedo et al., 2024), and the StarcoderData (Li et al., 2023b) corpus, as integrating code is known to boost general model performance (MA et al., 2024). For multimodal pre-training, we used public benchmarks like ImageNet (Russakovsky et al., 2015) and JourneyDB (Pan et al., 2023), complemented by a substantial internally collected dataset of 40 million images, which were captioned using the powerful Intern-VL model Chen et al. (2024). Following the methodology of Liquid (Wu et al., 2025b), we diversified our training data by randomly reversing 20% of the text-to-image pairs to serve as image-captioning tasks. The final pre-training data consists of 72B text tokens and 65B vision tokens.

Supervised Fine-Tuning (SFT). We further refined the model with 3B SFT tokens. For vision understanding, we employed MiniGemini (Li et al., 2024) and FineVision (HuggingFaceM4, 2025). To improve text understanding and general instruction-following, we utilized OpenOrca (Mukherjee et al., 2023). Additionally, to refine the quality of image generation, we leveraged Blip3o-60k (Chen et al., 2025a) and ShareGPT4o (Chen et al., 2023).

Benchmarks. Evaluation covered text-only, image generation, and multimodal understanding tasks. For text-only tasks, we employed ARC-Easy/Challenge (**ARC-E/ARC-C**) (Clark et al., 2018), WinoGrande (**WinoG**) (Sakaguchi et al., 2020), **BoolQ** (Clark et al., 2019), and **MMLU** (Hendrycks et al., 2021). For image generation, we used **GenEval** (Ghosh et al., 2023) and DPG-Bench (**DPG**) (Hu et al., 2024). For the GenEval benchmark, we followed Bagel (Deng et al., 2025) and employed an LLM to rewrite shorter prompts into more detailed ones to better assess instruction following. For multimodal understanding, we used SEEDBench (**SEED**) (Li et al., 2023a), **MME** (Fu et al., 2024), **POPE** (Li et al., 2023c), and MMBench (**MMB**) (Liu et al., 2024d).

Implementation Details. We conducted ablation studies on Qwen2.5-1.5B (Yang et al., 2024) and scaled to Qwen2.5-3B. We used the VQGAN tokenizer (Esser et al., 2021a) from Chameleon (Team,

Table 1: The text performance of Uni-X compared to other models.

Model	# Params.	ARC-E	ARC-C	WinoG	BoolQ	MMLU	Avg. ↑
Janus-Pro (Chen et al., 2025b)	7B	70.4	40.9	66.1	80.2	49.3	61.4
VILA-U (Wu et al., 2024b)	7B	51.6	34.0	57.3	70.6	25.5	47.8
Chameleon (Team, 2025)	7B	76.1	46.5	70.4	81.4	52.1	65.3
Liquid (Wu et al., 2025b)	7B	75.6	49.0	72.7	81.0	56.0	66.9
Uni-X	3B / 4.5B	79.0	47.9	68.9	82.2	57.6	67.1

Table 2: The image generation and multimodal understanding performance of Uni-X compared to other models. In the **# Params** column x/y, x and y represent the number of active parameters and the total parameters, respectively. † represents the model variant that performs semantic alignment. ‡ represents the rewriting of the prompt during evaluation. $^\heartsuit$ indicates that it has been trained on more image-text data.

Model	# Tokens	# Params.	GenEval	DPG	MME	POPE	MMB	SEED
Autoregressive meets Diffusion								
Bagel (Deng et al., 2025)	5.1T	7B / 14B	$88^\ddagger$	85.0	-	-	85.0	-
Bilp3o (Chen et al., 2025a)	-	4B / 9B	81	79.3	1,527.7	-	78.6	73.8
X-Omni (Geng et al., 2025)	~1T	10B / 20B	$83^\ddagger$	87.6	-	89.3	74.8	74.1
Show-o (Xie et al., 2024)	~500B	1.3B	68	-	1,097.2	80.0	-	-
Show-o† (Xie et al., 2024)	~500B	1.3B	69	-	1,232.9	84.5	-	-
Autoregressive w/ Semantic Encoder								
NextStep1 (Team et al., 2025)	~1T	14B	$73^\ddagger$	85.2	-	-	-	-
Janus-Pro (Chen et al., 2025b)	~300B	7B	80	84.1	-	87.4	79.2	72.1
VILA-U (Wu et al., 2024b)	-	7B	-	-	1,336.2	83.9	-	56.3
Liquid† (Wu et al., 2025b)	-	8B	-	-	1,448.0	83.2	-	-
Autoregressive w/o Semantic Encoder								
Chameleon (Team, 2025)	9.2T	34B	39	-	604.5	-	32.7	-
LWM (Liu et al., 2024a)	~500B	7B	47	-	-	75.2	-	-
EMU3 (Wang et al., 2024d)	-	8B	$66^\ddagger$	80.6	1,243.8	85.2	58.5	68.2
Liquid (Wu et al., 2025b)	~90B	7B	$68^\ddagger$	79.8	1,107.2	81.1	-	-
Uni-X	140B	3B / 4.5B	$82^\ddagger$	79.8	1,158.3	83.6	59.3	60.2
Uni-X$^\heartsuit$	240B	3B / 4.5B	$83^\ddagger$	80.3	1,228.2	84.6	62.7	59.8

2025) to encode 512×512 images into 32×32 discrete tokens. Our codebase is built upon the Liquid (Wu et al., 2025b) and HuggingFace Transformers (Wolf et al., 2019) libraries. Training was accelerated using Flash Attention 2 (Dao et al., 2022) and DeepSpeed ZeRO2 (Aminabadi et al., 2022). When generating images, we uniformly set the classifier-free guidance (CFG) to 4.0.

4.2 Results and Analysis

Scaling Experiment. In this experiment, we aim to demonstrate that Uni-X can scale effectively and is not limited to small-scale training data. We expand the dataset size to 140B total tokens for Uni-X, using Qwen2.5-3B as the base model, and extend the training duration to achieve improved performance. The evaluation is conducted against SOTA models, some of which have been trained on trillions of tokens. As presented in Table 1, Uni-X robustly maintains the strong language capabilities of its base model. With an average score of 67.1 across five text benchmarks, our 3B Uni-X model outperforms several larger 7B models. This demonstrates that our design successfully mitigates modality conflict without sacrificing performance on fundamental language understanding tasks.

For image generation, as detailed in Table 2, Uni-X achieves a strong score of 82 on GenEval, a result that surpasses many models with more parameters and underscores the effectiveness of

Table 3: Image edit results on ImgEdit-Bench. ♡ indicates that it has been trained on more image-text data.

Model	# Params.	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall↑
GPT-4o	-	4.61	4.33	2.90	4.35	3.66	4.57	4.93	3.96	4.89	4.20
ICEdit	12B	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
AnyEdit	4B	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit	4B	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
Step1X-Edit	12B	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
Bagel	7B / 14B	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
Uni-X♡	3B / 4.5B	3.57	3.18	2.06	3.94	3.82	3.38	4.21	3.16	3.63	3.44

Table 4: Performance and training efficiency comparison of different model architectures under identical training conditions. Training efficiency is measured by the number of tokens processed per second per GPU. ♠ indicates that the baseline has been adapted for our experimental setting (see Appendix A.2 for specific details); ◇ represents calculating the loss on the instruction part during the training of image-text data.

Model	#Params.	MMLU	GenEval	MMB	Avg. ↑	Efficiency ↑
Shared Transformer	1.5B	50.0	33.6	30.3	38.0	16,380
MoT♠ Deng et al. (2025)	1.5B / 3B	48.0	26.0	30.0	34.6	12,658
HardMoE	1.5B / 2.3B	50.3	42.8	30.7	41.3	14,657
UniFork♠ (Li et al., 2025c)	1.5B / 2.3B	50.1	12.4	25.9	29.5	15,481
Uni-X (9:5)◇	1.5B / 2.3B	48.5	34.8	29.8	37.7	15,642
Uni-X (9:5)	1.5B / 2.3B	50.1	43.3	31.5	41.6	15,595

our architecture in producing high-quality images. We also tested T2I-CompBench (Huang et al., 2025) and MSCOCO (Lin et al., 2015), as shown in Appendix A.3. Uni-X similarly exhibited strong performance with fewer parameters. Regarding vision understanding (Table 2), while Uni-X's scores are slightly lower than some state-of-the-art models, we observe a clear trend: models that incorporate an additional semantic image encoder, such as Janus-Pro (Chen et al., 2025b) and the semantically aligned variants of Liquid (Wu et al., 2025b) and Show-o (Xie et al., 2024), tend to achieve substantially higher performance on understanding benchmarks like MMBench and SEED.

In contrast, among models that do not rely on a separate semantic encoder, Uni-X's performance is commendable and holds its ground against strong competitors like EMU3 (Wang et al., 2024d). This suggests that our architecture effectively harnesses the inherent capabilities of the autoregressive framework for vision understanding. We speculate that the relatively weaker understanding performance might be partially caused by the insufficient utilization of the VQ tokenizer's codebook. We analyzed the token sequences encoded from 1 million images and found that, although there are 8,192 tokens available in the tokenizer, only ≈3,127 are being utilized. Meanwhile, EMU3 uses 4096 tokens to represent a 512×512 image, which provides more fine-grained information. However, this $4\times$ token count severely impacts its image generation speed, as shown in Appendix A.5.

For image editing, we conducted tests on ImgEdit (Ye et al., 2025) as shown in Table 3. Uni-X achieved better results than Bagel, even with less training data and fewer parameters than Bagel. This demonstrates that the high-level semantic unification of Uni-X enhances its image editing capabilities.

Identical Training Conditions. To validate the effectiveness of Uni-X, we conducted ablation experiments on a smaller dataset and a slightly reduced base model Qwen2.5-1.5B, due to resource constraints. To ensure consistency in performance comparisons, we limited the dataset to 28B tokens, of which 13.7B are vision tokens. The experiments were conducted using learning rate (LR) 5×10^{-5}, warmup ratio 0.03, and constant LR scheduler, with batch size 17,560 tokens per GPU.

The selected baselines include: (1) **Shared Transformer**, which continues multimodal pre-training based on Qwen2.5-1.5B; (2) **Mixture-of-Transformers (MoT)** (Deng et al., 2025), where prior work replicates an additional transformer to handle image generation tasks, while the original LLM backbone focuses on text-only and image understanding tasks. Under our experimental setup,

Figure 5: Qualitative examples of Uni-X image generation. The results highlight its ability to produce diverse, high-quality visuals that follow prompts with both creativity and fine-grained detail.

vision tokens are allocated to the duplicated transformer; (3) **Hard-Route MoE (HardMoE)**, which introduces a vision expert specifically for the vision modality, assigning vision tokens to this expert for computation guided by the vision mask; and (4) **UniFork** (Li et al., 2025c), which creates a task-specific deep branch for image generation. For all the baselines, we ignored the instruction during training to enhance cross-modal performance and ensure a fair comparison.

Results (Table 4) show that Uni-X achieves the best overall performance under consistent training conditions. Specifically, our Uni-X (9:5) configuration attains an average score of 41.6, significantly outperforming the standard baselines. While HardMoE is competitive, achieving a score of 41.3, Uni-X still holds a slight advantage. Moreover, HardMoE and UniFork are orthogonal and can be combined. In terms of training efficiency, although the baseline shared transformer is the fastest due to having the fewest parameters, Uni-X achieves a high throughput, which is considerably more efficient than the less performant MoT architecture. These findings confirm that Uni-X's design offers a more effective trade-off between performance and computational efficiency.

It is worth noting that the architectures of MoT and UniFork have been adapted to fit our VQ+AR setup to avoid discrepancies in efficiency between paradigms such as diffusion and AR+diffusion. Specific details can be found in Appendix A.2. A comparison of training efficiency across paradigms lies beyond the scope of this work and will be considered in future research.

Case Study. In Figure 5, we present a curated selection of images generated by Uni-X to qualitatively assess its capabilities. Despite the relatively limited number of tokens used during training, the model demonstrates a strong ability to produce clear, aesthetically pleasing images that exhibit robust instruction-following capabilities. The examples showcase Uni-X's versatility in handling a wide range of creative and complex prompts. For instance, the model can generate imaginative fantasy scenes, such as a gigantic library floating above the clouds, and surreal compositions, like a realistic elephant walking on the ocean floor. Furthermore, Uni-X successfully adheres to specific artistic style requests, as seen in the detailed anime-style portrait, and renders fine details with high fidelity, exemplified by the intricate feather patterns of the owl. These case studies collectively verify that Uni-X can effectively translate complex textual descriptions into high-quality visual outputs. The specific prompts used for these generations are provided in Appendix A.1.

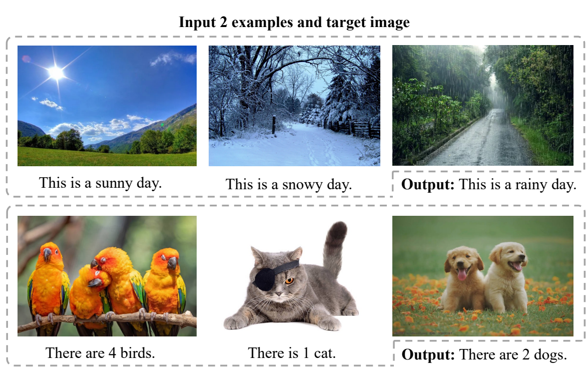

Figure 6: Demonstration of Uni-X's in-context learning. The model follows few-shot examples to perform tasks such as image description (1st line) and object counting (2nd line).

In-Context Learning. Although Uni-X was not explicitly trained on interleaved multimodal data, we conducted an evaluation to assess its emergent in-context learning (ICL) capabilities. As illustrated in Figure 6, the model was presented with few-shot examples, where several image-text pairs were provided as context before a final query image was presented without its corresponding description.

The results demonstrate that Uni-X can successfully interpret the contextual examples and apply the learned pattern to the target image. For instance, in the top row of Figure 6, the model correctly identifies the weather in the target image as a "rainy day," adhering to the simple descriptive format ("This is a... day.") established by the preceding examples. Also, Uni-X exhibits the ability to perform more reasoning tasks such as object counting. This suggests that the model is not merely mimicking sentence structure but is performing cross-modal reasoning at a semantic level.

Ignore Instruction in Training. Ignoring the loss of the instruction part during training is a common technique in supervised fine-tuning. However, its role in pretraining is rarely emphasized. Following Liquid (Wu et al., 2025b), we applied the same "ignore instruction" strategy during pretraining.

Specifically, no loss mask was applied for pure text data. For text-image pairs, in text-to-image tasks, the loss calculation excluded the text instruction tokens. Similarly, for image captioning tasks, the loss corresponding to the image tokens was masked. As demonstrated in our experimental results Table 4, this approach significantly enhanced the model's capability to generate images.

We believe there might be several reasons for this: 1) This mask forces the model to learn the relationship between the two modalities rather than relying on the prior distribution of images, thereby enhancing its instruction-following capability. 2) It serves as a form of loss regularization. For text-image pair data, the number of image tokens is fixed at 1024, while the average number of text tokens is around 120. By masking, we ensure that the gradient magnitude generated by the loss is only dependent on the reverse ratio we set.

Number of Separated Layers. We investigate how the number and distribution of separated layers affect performance (Table 5). Varying the total number of separated layers produces an n-shaped trend: more separation improves modality-specific low-level processing, but too many reduce shared middle layers, weakening semantic fusion and cross-modal reasoning. The best overall performance is achieved with 14 separated layers. We then examine shallow-deep ratios under this setting. A 9:5 split (slightly more shallow than deep layers) performs best, indicating that early processing of low-level features, where text and vision differ most, benefits more from modality-specific capacity than the final generation stage. These results provide strong empirical support for the Uni-X design. We also explored text layers and vision layers with different numbers of separate layers, and the results are shown in Appendix A.6.

Table 5: Performance comparison of different Uni-X configurations. Here, $x : y$ denotes the number of shallow separated layers x and deep separated layers y, respectively. The total number of layers is $n = 28$. The split points are x and $n - y$, respectively.

Configuration	MMLU	GenEval	MMB	Avg. ↑
Uni-X (3:3)	48.7	37.3	30.7	38.9
Uni-X (7:7)	49.6	41.3	29.4	40.1
Uni-X (11:11)	49.7	37.5	32.1	39.8
Uni-X (3:11)	50.0	32.9	31.0	38.0
Uni-X (5:9)	50.1	39.2	28.0	39.1
Uni-X (9:5)	50.1	43.3	31.5	41.6
Uni-X (11:3)	49.8	25.1	31.9	35.6

5 Conclusions

In this work, we identified gradient conflicts as a fundamental limitation of shared AR UMMs, particularly in the shallow and deep layers where vision and text exhibit highly divergent low-level statistics. To address this challenge, we proposed Uni-X, a two-end-separated, middle-shared architecture that explicitly aligns model structure with modality characteristics. By isolating low-level processing into modality-specific branches while maintaining a shared semantic core for high-level fusion, Uni-X effectively mitigates inter-modal conflicts without adding architectural complexity. Extensive experiments show that this X-shaped design allows a 3B-parameter Uni-X model to deliver performance competitive with much larger 7B UMMs across diverse multimodal benchmarks. These findings establish Uni-X as both a scalable and parameter-efficient foundation, paving the way for future research in unified multimodal modeling.

Ethics Statement

This research aims to advance the field of artificial intelligence, particularly in the area of Unified Multimodal Models. We recognize that, like other powerful generative models, the technologies proposed in this study also carry potential risks of misuse, such as the creation of misinformation, biased, or harmful content. Our primary objective is to explore architectural efficiency to build more powerful and scalable models, and we believe this will make a valuable contribution to science.

The datasets used to train and fine-tune our models are primarily publicly available and widely used benchmark datasets in the academic community. For any internally collected data, we have ensured that its acquisition and processing adhere to principles of responsibility. We have not specifically filtered web-based datasets for bias, and therefore, the model may reflect social biases present in the data. We encourage responsible downstream use and further research into mitigating the potential negative impacts of generative models. Our work is intended solely for research purposes and is shared with the community to foster innovation and deepen understanding.

Reproducibility Statement

We are committed to ensuring the reproducibility of our research. To this end, we provide comprehensive details throughout the paper and in the appendix.

Code. The source code for our Uni-X model architecture, training, and evaluation is made available at the following anonymous repository: `https://anonymous.4open.science/r/Uni-X-Code-E5CD`.

Architecture and Implementation. The detailed architecture of Uni-X is described in Section 3.2. Implementation details, including the base models used (Qwen2.5-1.5B and Qwen2.5-3B), the VQGAN tokenizer, and software dependencies are provided in Section 4.1. Details on our baseline implementations are available in Appendix A.2.

Datasets and Evaluation. All datasets used for pre-training and supervised fine-tuning are listed in Section 4.1. The evaluation benchmarks for understanding and generation tasks are also detailed in the same section.

Hyperparameters. Key hyperparameters for our main ablation study, including learning rate, batch size, and scheduler details, are specified in Section 4.3 to ensure a fair comparison.

We believe that the combination of our provided code, detailed architectural descriptions, dataset lists, and specific hyperparameters will enable the community to replicate our findings and build upon our work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. DeepSpeed-inference: enabling efficient inference of transformer models at unprecedented scale. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC'22)*, pp. 1–15, 2022.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024. URL `https://arxiv.org/abs/2407.07726`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. BLIP3-o: A Family of Fully Open Unified Multimodal Models-Architecture, Training and Dataset, May 2025a. URL `http://arxiv.org/abs/2505.09568`. arXiv:2505.09568 [cs].

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling, January 2025b. URL `http://arxiv.org/abs/2501.17811`. arXiv:2501.17811 [cs].

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL `https://arxiv.org/abs/1803.05457`.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 16344–16359, 2022.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging Properties in Unified Multimodal Pretraining, 2025. URL `https://arxiv.org/abs/2505.14683`. Version Number: 2.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021a.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021b. URL `https://arxiv.org/abs/2012.09841`.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation, 2025. URL `https://arxiv.org/abs/2404.14396`.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, Linus, Di Wang, and Jie Jiang. X-Omni: Reinforcement Learning Makes Discrete Autoregressive Image Generative Models Great Again, July 2025. URL `http://arxiv.org/abs/2507.22058`. arXiv:2507.22058 [cs].

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer Feed-Forward Layers Are Key-Value Memories, September 2021. URL http://arxiv.org/abs/2012.14913. arXiv:2012.14913 [cs].

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. URL https://arxiv.org/abs/2310.11513.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024. URL https://arxiv.org/abs/2403.05135.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation, 2025. URL https://arxiv.org/abs/2307.06350.

HuggingFaceM4. Finevision dataset. https://huggingface.co/datasets/HuggingFaceM4/FineVision, 2025. https://huggingface.co/datasets/HuggingFaceM4/FineVision.

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding, 2025. URL https://arxiv.org/abs/2504.04423.

Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization, March 2024. URL http://arxiv.org/abs/2309.04669. arXiv:2309.04669 [cs].

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, Chongyan Zhu, Xiaoyi Ren, Chao Li, Yifan Ye, Peng Liu, Lihuan Zhang, Hanshu Yan, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model, 2025a. URL https://arxiv.org/abs/2410.05993.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2025b. URL https://arxiv.org/abs/2406.11794.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you!, 2023b. URL https://arxiv.org/abs/2305.06161.

Teng Li, Quanfeng Lu, Lirui Zhao, Hao Li, Xizhou Zhu, Yu Qiao, Jun Zhang, and Wenqi Shao. UniFork: Exploring Modality Alignment for Unified Multimodal Understanding and Generation, June 2025c. URL http://arxiv.org/abs/2506.17202. arXiv:2506.17202 [cs].

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models, March 2024. URL http://arxiv.org/abs/2403.18814. arXiv:2403.18814 [cs].

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023c.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An Omni Foundation Model for Interleaved Multi-Modal Generation, May 2025. URL http://arxiv.org/abs/2505.05472. arXiv:2505.05472 [cs].

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024c.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024d. URL https://arxiv.org/abs/2307.06281.

Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help llms reasoning?, 2023. URL https://arxiv.org/abs/2309.16298.

YINGWEI MA, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help LLMs reasoning? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KIPJKST4gw.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4, 2023. URL https://arxiv.org/abs/2306.02707.

Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 30811–30849, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL `https://arxiv.org/abs/1409.0575`.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 8732–8740, 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL `https://arxiv.org/abs/2402.03300`.

Guangyuan Shi, Qimai Li, Wenlong Zhang, Jiaxin Chen, and Xiao-Ming Wu. Recon: Reducing conflicting gradients from the root for multi-task learning, 2023. URL `https://arxiv.org/abs/2302.11289`.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. LMFusion: Adapting Pretrained Language Models for Multimodal Generation, February 2025. URL `http://arxiv.org/abs/2412.15188`. arXiv:2412.15188 [cs].

Qi Sun, Marc Pickett, Aakash Kumar Nain, and Llion Jones. Transformer layers as painters, 2025. URL `https://arxiv.org/abs/2407.09298`.

Chameleon Team. Chameleon: Mixed-Modal Early-Fusion Foundation Models, March 2025. URL `http://arxiv.org/abs/2405.09818`. arXiv:2405.09818 [cs].

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024a.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024b.

NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, Hongyu Zhou, Kenkun Liu, Ailin Huang, Bin Wang, Changxin Miao, Deshan Sun, En Yu, Fukun Yin, Gang Yu, Hao Nie, Haoran Lv, Hanpeng Hu, Jia Wang, Jian Zhou, Jianjian Sun, Kaijun Tan, Kang An, Kangheng Lin, Liang Zhao, Mei Chen, Peng Xing, Rui Wang, Shiyu Liu, Shutao Xia, Tianhao You, Wei Ji, Xianfang Zeng, Xin Han, Xuelin Zhang, Yana Wei, Yanming Xu, Yimin Jiang, Yingming Wang, Yu Zhou, Yucheng Han, Ziyang Meng, Binxing Jiao, Daxin Jiang, Xiangyu Zhang, and Yibo Zhu. NextStep-1: Toward Autoregressive Image Generation with Continuous Tokens at Scale, August 2025. URL `http://arxiv.org/abs/2508.10711`. arXiv:2508.10711 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL `https://arxiv.org/abs/2502.14786`.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL https://arxiv.org/abs/1711.00937.

Liangdong Wang, Bo-Wen Zhang, Chengwei Wu, Hanyu Zhao, Xiaofeng Shi, Shuhao Gu, Jijie Li, Quanyue Ma, TengFei Pan, and Guang Liu. Cci3.0-hq: a large-scale chinese dataset of high quality designed for pre-training large language models, 2024a. URL https://arxiv.org/abs/2410.18505.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b. URL https://arxiv.org/abs/2409.12191.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-Token Prediction is All You Need, September 2024c. URL http://arxiv.org/abs/2409.18869. arXiv:2409.18869 [cs].

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024d.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation, October 2024a. URL http://arxiv.org/abs/2410.13848. arXiv:2410.13848 [cs].

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Exploration to Advanced Multimodal Generation, June 2025a. URL http://arxiv.org/abs/2506.18871. arXiv:2506.18871 [cs].

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language Models are Scalable and Unified Multi-modal Generators, April 2025b. URL http://arxiv.org/abs/2412.04332. arXiv:2412.04332 [cs].

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024b.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved Native Unified Multimodal Models, June 2025. URL http://arxiv.org/abs/2506.15564. arXiv:2506.15564 [cs].

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark, 2025. URL https://arxiv.org/abs/2505.20275.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, 2020. URL `https://arxiv.org/abs/2001.06782`.

Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

A Appendix

A.1 Prompts of Image Generation

Table 6 lists the prompts corresponding to the generated images shown in Figure 5. The prompts are presented in the same order as the images: left to right, top to bottom. These examples highlight the diversity of tasks, ranging from descriptive captions to creative scene generation.

Table 6: Prompts used for the image generation examples shown in Figure 5.

No.	Prompt
1	A gigantic library floats above the clouds, its appearance resembling a suspended castle. Every book emits a faint glow and drifts through the air with the gentle breeze.
2	A highly realistic close-up photo featuring a beautiful 35-year-old red-haired woman, writing in her diary on her balcony. She is dressed in warm yet stylish clothing.
3	A happy snowman.
4	A woman and her little lion taking a selfie on the grassland.
5	A beautiful owl with sleek feathers and lively eyes, its round head adorned with two furry ears. The elegant pattern is formed by the interweaving of snow-white down and deep brown flight feathers, making it appear both stunning and endearing.
6	A clearing in a deep, mysterious forest, with a mirror-like pond at its center, the water reflecting a night sky filled with the Milky Way.
7	A handsome 24-year-old boy stands in the center, with a sky-colored background. He is wearing glasses, and the art style is very detailed, in anime style.
8	A realistic photo of an elephant walking on the ocean floor.
9	An elegant and charming lady whose hair is entirely made up of blooming flowers, resembling a masterpiece of nature. The flowers are of various types, possibly including delicate roses, fresh daisies, vibrant sunflowers, or other colorful blossoms.
10	A magnificent landscape photo depicting the northern lights dancing above the snow-capped mountain ranges in Iceland.

A.2 Baseline Implementation Details

To ensure fair comparisons, we adapt baseline methods to the VQ+AR setting used in our study. For Mixture-of-Transformers (MoT) (Deng et al., 2025; Shi et al., 2025; Liao et al., 2025), the duplicated transformer is originally designed for image generation through diffusion. To remove the influence of diffusion and isolate architectural effects, we reconfigure the duplicated transformer to operate directly on image tokens. In this setup, the `qkv` sequences from the two transformers are concatenated within the attention module, allowing the model to incorporate visual information for both understanding and generation tasks. As a result, the MoT results reported in this paper reflect its effectiveness strictly within the VQ+AR paradigm, eliminating confounding factors introduced by diffusion-based processes.

Table 7: The T2I-CompBench and MSCOCO performance of Uni-X. ♡ indicates that it has been trained on more image-text data.

Model	# Params.	T2I-Color	T2I-Shape	T2I-Texture	T2I-Avg. ↑	MSCOCO CLIP-T
SDXL	3.5B	63.7	54.1	56.4	58.1	-
Janus	1.3B	75.5	47.7	62.1	61.8	-
Liquid	7B	71.5	52.3	65.1	63.0	30.7
EMU3	8B	61.1	47.3	61.9	56.8	31.3
UniToken	7B	71.2	51.8	66.7	63.2	-
Uni-X♡	**3B / 4.5B**	**76.5**	**56.3**	**67.1**	**66.6**	**31.8**

Table 8: Average gradient conflict between different domain data. Higher values indicate a higher degree of conflict.

Model	Code vs. Math	Code vs. Wiki	Math vs. Wiki
Qwen2.5-1.5B	0.158	0.330	0.262
Qwen2.5-3B	0.130	0.382	0.294
Qwen2.5-Coder-3B	0.182	0.317	0.275
Qwen2.5-7B	0.153	0.263	0.240
Llama3.2-3B	0.297	0.351	0.360

A.3 More Evaluation Results on Image Generation Benchmark

We conducted tests on the T2I-CompBench (Huang et al., 2025) and MSCOCO (Lin et al., 2015). Part of the results were excerpted from UniToken (Jiao et al., 2025). As shown in Table 7, Uni-X surpassed the recent strong autoregressive models EMU3 and Liquid in the newly added image generation benchmark. Uni-X also achieved better results than UniToken, which includes semantic information.

A.4 Gradient Conflict Analysis

Analysis on Mainstream Models. We further demonstrate the effectiveness of the current gradient conflict metric through experiments. We conduct a quantitative analysis on mainstream models such as Qwen and Llama, as shown in Table 8. For each dataset, we utilized a total of 2M tokens (accumulated over 60 batches) to compute gradients, to ensure minimal gradient noise.

All models in Table 8 exhibit a consistent pattern: the gradient conflict between Code vs. Math is strictly lower than for both Code vs. Wiki and Math vs. Wiki. It is well-established in LLM pre-training that Code and Math tasks often mutually enhance each other (Shao et al., 2024; Ma et al., 2023). This phenomenon is precisely reflected in our gradient conflict analysis.

The relatively high gradient similarity (low conflict) between these two tasks implies that improvements in Code performance can drive improvements in Math performance. We further verified this in Table 9. Qwen2.5-Coder-3B, which was fine-tuned from Qwen2.5-3B to specifically enhance coding capabilities, simultaneously achieved a substantial improvement in Math performance. This validates our hypothesis that lower gradient conflict correlates with positive transfer between modalities/domains.

Analysis on Other Modules. In Section 3.1 of the main text, we analyzed gradient conflicts in the down-projection weights of the Feed-Forward Network (FFN). To develop a more complete picture and confirm that this issue is not confined to a single component, we extend our analysis to additional modules of the transformer. In particular, we examine gradient conflicts in the output projection weights (`O_PROJ`) and value projection weights (`V_PROJ`) of the self-attention mechanism, both of which play critical roles in multimodal representation learning.

Using the same methodology for conflict measurement, Figures 7 and 8 reveal a consistent trend with that observed in the FFN layers. The modality-shared transformer exhibits severe gradient conflicts

Table 9: Domain performance of Qwen2.5-3B and Qwen2.5-Coder-3B under zero-shot settings.

Model	**HumanEval (Code)**	**GSM8K (Math)**	**MMLU (Wiki)**
Qwen2.5-3B	39.0	6.0	65.0
Qwen2.5-Coder-3B	45.7	26.1	60.8

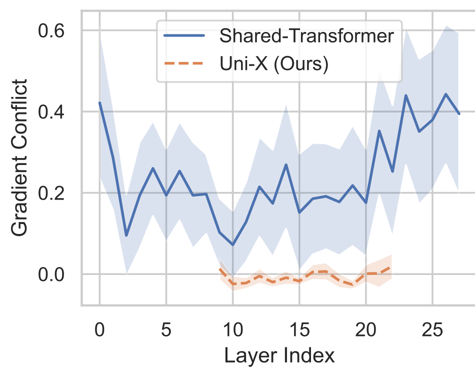

Figure 7: An analysis of gradient conflict in attention of out projection weights.

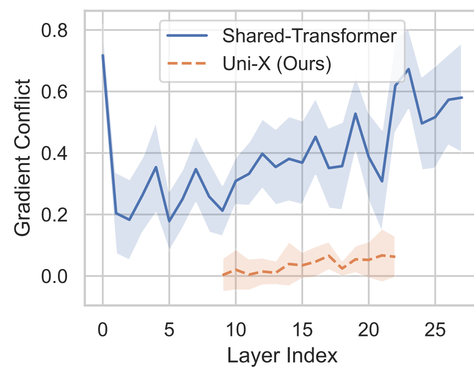

Figure 8: An analysis of gradient conflict in attention of value projection weights.

in the shallow and deep layers of both the attention output and value projection weights, with only partial alleviation in the middle layers. In contrast, Uni-X effectively addresses these issues: (i) modality-specific layers at both ends prevent conflicts in low-level processing and output stages, and (ii) the shared middle block further reduces residual conflicts by leveraging semantic alignment.

These results strengthen our hypothesis that gradient conflict stems from the intrinsic statistical mismatch between vision and text, and they demonstrate that Uni-X's two-end-separated, middle-shared design offers a robust and generalizable solution across multiple transformer components.

A.5 Inference Efficiency

We have conducted a comprehensive evaluation of inference efficiency on an H800 PCIe (350W) GPU. As shown in Table 10, Uni-X demonstrates superior throughput compared to standard autoregressive baselines.

Uni-X achieves high throughput (910.2 tokens/s) even compared to the original Qwen2.5-3B (975.2 tokens/s), despite the architectural changes and a higher number of parameters (4.5B vs 3B). This efficiency gain stems from the computational complexity of the attention mechanism in the separated layers.

Theoretically, the computational cost of the Uni-X architecture is lower, and the current inference speed still has a slight gap because the current code has not been fully optimized. In the separated layers, a sequence of length n is effectively partitioned into vision tokens of length a and text tokens of length b (where $a + b = n$). Since the self-attention complexity is $O(n^2)$, and the separated layers enforce strict modality isolation, the complexity reduces to proportional to $a^2 + b^2$. Since $a^2 + b^2 < (a+b)^2 = n^2$, the computational cost for attention in these specific layers is strictly lower than in a fully shared transformer, leading to the observed speedup.

A.6 Ablation Study on Ratio between Text and Vision.

We conducted experiments maintaining the same hyperparameters and training volume as in Table 5, and the results are shown in Table 11. We continued to use Qwen2.5-1.5B with a total of 28 layers as the base model. The number of vision layers directly affects the performance related to image understanding and generation. Surprisingly, reducing the number of vision layers also decreases pure text performance. This may be because the shared layers in the middle have to process more

Table 10: Inference throughput comparison. Settings: batch size 48, input length ≈1,200 tokens, outputting one image.

Model	# Params.	Throughput ↑	
		Tokens/s	Images/min
Shared Transformer (Qwen2.5-3B)	3B	975.2	-
Liquid	7B	182.0	10.6
EMU3	8B	199.0	2.9
Uni-X	3B / 4.5B	910.2	**53.3**

Table 11: Ratio between t-layers and v-layers within the separated layers.

Configuration	MMLU	GenEval	MMB	Avg. ↑
14:8	48.2	37.8	26.1	37.4
14:14	49.6	41.3	29.4	40.1
14:20	50.1	42.6	31.0	41.2

low-level vision information, thereby leading to a decline in pure text capability. This experimental result also proves the effectiveness of our proposed architecture from another perspective.

A.7 Use of Large Language Models

Large Language Models (LLMs) were used solely as writing aids during manuscript preparation. Their role was limited to language polishing, improving grammar, clarity, and readability, without influencing the conceptual design, experimental methodology, or analytical findings. All research ideas, model designs, and experimental results are the original contributions of the authors.